

POSTPROCESSING TECHNIQUES FOR VOICE PITCH TRACKERS

Bruce G. Secrest and George R. Doddington

Texas Instruments Inc.
Central Research Laboratories
13510 N. Central Expressway
P.O.Box 226015, MS 238
Dallas, Texas 75266, USA

ABSTRACT

The concepts of dynamic programming and pattern matching are used in a postprocessing technique for voice pitch tracking. The dynamic programming is used to achieve a smooth pitch contour based on information in the current and previous frames. A voicing decision is made based on matching the smooth contour from the dynamic programming with a set of reference templates which are representative of voiced and unvoiced speech. Three postprocessing techniques; the dynamic programming technique, nonlinear smoothing and median smoothing are appended to three standard pitch tracking algorithms. A comparative evaluation of these postprocessing techniques is performed over a data base of 58 male and female speakers ranging from 6 to 87 years of age. It is shown that significant improvement in the performance of the pitch tracking algorithms can be obtained with the dynamic programming postprocessing.

INTRODUCTION

Most pitch tracking algorithms fail because of anomalous (aperiodic) behavior of the vocal cord vibrations of the speaker. This behavior is attributed to such things as start-up problems at the beginning of voiced segments or by low speech effort at the end of a phrase. The anomalies can manifest themselves as irregular pitch periods or as regions where every other pitch pulse varies widely in amplitude. If a person is visually extracting a pitch contour, he will look to nearby areas of relatively periodic speech to extend the pitch into the region of difficulty. This same type of approach can be utilized in pitch tracking algorithms by introducing a postprocessing technique in which the decision of choosing a pitch period for a frame of data is delayed until several more frames have been evaluated.

The postprocessing technique utilizing dynamic programming requires that a set of candidate pitch periods be provided for each frame of data along with a measure of the goodness of each candidate[1]. For example, the lags of the normalized correlation function peaks can be used as the pitch candidates, and the magnitude of the corresponding peaks can be used as the goodness of the candidate. Dynamic programming is utilized to construct a series of candidate pitch tracks from frame to frame with a cumulative penalty assigned

to each track. A smooth pitch contour is obtained by choosing the track with the minimum penalty. A voicing decision for each frame is made by comparing the goodness values associated with the smooth contour with the reference templates shown in Figure 1. A scanning error is obtained for each template by calculating the squared error between the template and the goodness values of the optimum pitch track. The voiced/unvoiced decision is made by comparing the scanning errors against fixed thresholds[1].

Three postprocessing techniques: dynamic programming, nonlinear smoothing [2] and median smoothing are appended to the Gold Rabiner[3], cepstral[4], and the correlation pitch tracking algorithms. A comparative evaluation of these postprocessing techniques is performed over a data base of 58 male and female speakers. Objective test measures were used to compare the pitch contours from the postprocessing techniques with a hand edited 'reference' pitch contour. The objective measures used are shown to be highly correlated with subjective listening tests conducted over the same data base.

EXPERIMENT

To evaluate the performance of the postprocessing techniques, data were collected from 58 speakers, 32 males and 26 females ranging in age from 6 to 87 years. The subjects were prompted with one of 11 phrases.

A "reference" pitch contour was generated for each of the 58 tokens of this test data set by first extracting a pitch track for each token by using the cepstral pitch tracking algorithm; then the original waveform, a synthesized waveform and the pitch contour were displayed on a graphics terminal. By visually comparing the waveforms and listening to the original and synthesized speech, a decision was made interactively for the pitch period in each 10 millisecond frame of speech.

Subjective listening tests were conducted over the test data set by 15 listeners. The following five pitch contours were generated for each phrase of the data set:

1. "Reference" pitch contour (hand edited starting with cepstral pitch track).
2. Cepstral pitch track.
3. Gold Rabiner pitch track.
4. Correlation pitch track (basic correlation

pre-processing with dynamic programming postprocessing).

5. Edited correlation (obtained from correlation pitch track with hand editing to remove major pitch and voicing errors).

The subjective tests consisted of a series of A/B comparisons in which the listener showed a preference for token A, B, or neither. The tokens were synthesized speech generated by a 16 pole LPC model with a 10 millisecond frame period and pitch contours A or B used for the excitation. The two pitch contours for each comparison were obtained by choosing two contours at a time from the above five choices for each phrase.

Since objective testing is much easier than subjective testing, objective distortion measures were also used to evaluate the above pitch contours and the results were correlated with the results of the subjective listening tests over the same test set. The assertion is that if there is a high correlation between the objective and subjective results over the same test set, then the objective distortion measures can be used with some degree of confidence in further testing of similar distortions over the same test set. The objective measures were weighted distortion errors obtained relative to the "reference" pitch contour. The weighted distortion error, $E_{c,p}$ was obtained for each contour, c , and phrase, p , by summing for each frame in phrase p the following errors which occurred when compared with the "reference" contour:

$$\begin{aligned} PPE_i &= E_i / E_{\max} (|F_i - FR_i| / FR_i)^2 FR_i / 500 \\ \text{Voiced/Unvoiced error:} \\ VUVE_E &= E_i / E_{\max} (FR_i / 500) \\ VUVE_I &= E_i / E_{\max} (1 + FR_i / 500) \\ \text{Unvoiced/Voiced error:} \\ UVVE_E &= E_i / E_{\max} (F_i / 500) \\ UVVE_I &= E_i / E_{\max} (1 + F_i / 500) \end{aligned}$$

where PPE_i is the pitch period error, $VUVE_I$ is the voiced-unvoiced error in the interior of a reference voiced segment, $VUVE_E$ is the voiced-unvoiced error at the ends of a reference voiced segment, $UVVE_I$ is the unvoiced-voiced error in the interior of a test voiced segment and $UVVE_E$ is the unvoiced-voiced error at the ends of the test voiced segment, E_i is the RMS energy in the i 'th frame, $E_{\max}$ is the maximum RMS energy in the phrase, F_i is the pitch frequency in the i 'th frame for the test contour, FR_i is the reference pitch frequency in the i 'th frame, and 500 Hz is the maximum pitch frequency considered.

Because the subjective and objective test results could not be compared directly, functions of the test results were used for purposes of evaluation. Two such functions were chosen for the subjective tests and also two functions were chosen for the objective tests as given below.

First function.

For this first case, the function of the objective distortion error is given by the following quantities which express the preference of one pitch tracker over another:

$$O^1(1) = \sum_{j=1}^{58} (E_{2,j} - E_{1,j})$$

$$O^1(2) = \sum_{j=1}^{58} (E_{3,j} - E_{1,j})$$

$$O^1(3) = \sum_{j=1}^{58} (E_{4,j} - E_{1,j})$$

$$O^1(4) = \sum_{j=1}^{58} (E_{3,j} - E_{2,j})$$

$$O^1(5) = \sum_{j=1}^{58} (E_{4,j} - E_{2,j})$$

$$O^1(6) = \sum_{j=1}^{58} (E_{4,j} - E_{3,j})$$

where the $E_{c,j}$ are the above mentioned objective distortion errors of the c 'th pitch contour relative to the reference contour for the j 'th phrase. Here the c index corresponds to the following pitch trackers:

- $c = 1$ - correlation pitch track
- $c = 2$ - Gold-Rabiner
- $c = 3$ - cepstral
- $c = 4$ - edited correlation.

Therefore $O^1(1)$ would be the preference of the correlation pitch track relative to the Gold-Rabiner pitch track averaged over all 58 phrases.

Similarly the subjective function, $S^1(i), i=1,6$ is the corresponding preference shown by the listeners in the subjective listening test. That is, $S^1(1)$ is the total number of times (over all the phrases and all the listeners) the correlation pitch contour was preferred over the Gold-Rabiner contour minus the number of times the Gold-Rabiner contour was preferred over the correlation contour.

Second function.

The second function of the objective distortion errors to be considered is given by the following expression:

$$O^2(i) = \sum_{j=1}^{58} E_{i,j} ; i=1,2..4$$

where $E_{i,j}$ is again the objective distortion error of the i 'th pitch contour relative to the "reference" contour for the j 'th phrase. The corresponding function of the subjective test results is obtained by first calculating the total number of times a pitch contour is preferred by each listener. The subjective function is then

given by:

$$s^2(i) = \frac{\mu_R - \mu_i}{\mu_R}$$

where μ_R is the mean of the listeners preference for the "reference" contour and μ_i is the mean of the listeners preference for the i 'th contour.

The correlation coefficients for the two functions are given by:

$$\rho_j = \frac{\sum_{i=1}^M (o^j(i) - \bar{o}^j)(s^j(i) - \bar{s}^j)}{\left[\sum_{i=1}^M (o^j(i) - \bar{o}^j)^2 \right]^{1/2} \left[\sum_{i=1}^M (s^j(i) - \bar{s}^j)^2 \right]^{1/2}}$$

RESULTS

To establish the methods of evaluation, the results of the subjective listening test and their correlation with the objective measures are presented. Then the results of testing the various postprocessing techniques are given along with a discussion of these results.

A total of 15 listeners completed the subjective listening tests. The five contours shown above were compared on the test data set. The results of this test are given in Figures 2 and 3. Figure 2 shows the results of the first function above showing the preference of one pitch tracker over another having been averaged over all 58 phrases and the 15 listeners. Similarly, Figure 3 shows the results of applying the second function to the subjective test results, where the mean was obtained from each listener's preference for the particular pitch contour over all 58 phrases.

The result of comparing the last four contours above against the "reference" contour over the test data set is given using objective measures. The results of using the first function above are given in Figure 4 and the results for the second function are given in Figure 5.

Comparing the subjective listening test results with the objective test results for the first function gives a correlation coefficient $\rho_1=0.99$.

Similarly the results for the second functions of the objective and subjective test results gives the correlation coefficient $\rho_2=1.0$.

It can be seen that the correlations are high for both the functions. This is partly due to smoothing of the data because of averaging, but it is reasonable to assume that other pitch contours generated for this same data set would have similar strong correlations. Therefore the evaluations of the postprocessing techniques were performed with the objective testing methods.

The results of objective evaluation of the postprocessing techniques as implemented on the basic pitch tracking algorithms are given in figures 6 and 7.

CONCLUSIONS

It has been shown that significant improvement in the performance of three basic pitch tracking algorithms can be accomplished by using a dynamic programming postprocessing technique and that some improvement is obtained by using nonlinear smoothing and median smoothing postprocessing techniques.

ACKNOWLEDGEMENT

Portions of this work were supported by the Defense Advanced Research Projects Agency under contract number N00173-79-C-0224.

REFERENCES

- [1] B.G.Secretst and G.R.Doddington, "Evaluation of Postprocessing Techniques for Voice Pitch Trackers", to be submitted for publication.
- [2] L.R. Rabiner, M.R. Sambur and C.E.Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," IEEE trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp.552, Dec. 1975.
- [3] L.R. Rabiner and B. Gold, Theory and Application of Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, 1975.
- [4] ILS Application Note 1: Speech Analysis and Synthesis, Signal Technology, Inc. 15 W. DeLaGuerra Street, Santa Barbara, CA. 1979.

P	= { .5, .5, .5, .5 }	(Unvoiced)
UV		
P	= { .9, .9, .9, .9 }	(Voiced)
V		
P	= { .8, .8, .5, .5 }	(Voiced to Unvoiced transition)
VUV		
P	= { .5, .5, .8, .8 }	(Unvoiced to Voiced transition)
UVV		

Figure 1: Reference Voicing Patterns

		contour		
j\i		2	3	4
c	i			
o	1	311	173	-26
n	2		-108	-338
t	3			-232

Preference of contour i over contour j where:

- 1 - - correlation
- 2 - - Gold Rabiner
- 3 - - cepstral
- 4 - - edited correlation

Figure 2: Subjective test results using function S¹.

$$S^2(1) = \frac{\mu_R - \mu_1}{\mu_R} = .150$$

$$S^2(2) = \frac{\mu_R - \mu_2}{\mu_R} = .644$$

$$S^3(3) = \frac{\mu_R - \mu_3}{\mu_R} = .426$$

$$S^2(4) = \frac{\mu_R - \mu_4}{\mu_R} = .045$$

Preference of contour i relative to reference contour by comparing mean of listeners responses to contours:

- R - - Reference
- 1 - - correlation
- 2 - - Gold Rabiner
- 3 - - cepstral
- 4 - - edited correlation

Figure 3: Subjective test results using function S².

		contour		
j\i		2	3	4
c	i			
o	1	44	22	-11
n	2		-22	-56
t	3			-33

Objective preference of contour i over contour j where:

- 1 - - correlation
- 2 - - Gold Rabiner
- 3 - - cepstral
- 4 - - edited correlation

Figure 4: Evaluation results for function O¹ on objective test.

contour	pitch error	V-UV error	UV-V error	Total error
Edited correlation	.71	.15	.27	.58
Correlation	.72	.86	.39	1.60
Cepstral	.59	1.64	1.93	3.91
Gold Rabiner	.97	4.48	1.03	5.89

Figure 5: Evaluation results for function O² on objective test.

		CONTOUR												
j \ i		2	3	4	5	6	7	8	9	10	11	12	13	
C O U T O U R	1													
	2	11	.7	11	10	30	34	44	6	22	21	21	29	
	3		-4	0	0	19	23	33	-4	11	10	10	18	
	4			4	3	23	27	37	0	15	14	14	22	
	5				0	19	23	33	-4	10	10	10	18	
	6					20	23	33	-4	11	10	11	18	
	7						3	13	-24	-8	-9	-8	-1	
	8							9	-27	-12	-13	-12	-5	
R	9								-37	-22	-22	-22	-15	
	10									15	14	15	22	
	11										0	0	7	
	12											0	7	
	13												7	

Objective preference of contour i over contour j where:

- 1 - - correlation with dynamic programming
- 2 - - correlation with nonlinear smoothing
- 3 - - correlation with 3-pt median smoothing
- 4 - - correlation without postprocessing
- 5 - - Gold Rabiner with dynamic programming
- 6 - - Gold Rabiner with nonlinear smoothing
- 7 - - Gold Rabiner with 3 pt median smoothing
- 8 - - Gold Rabiner without postprocessing
- 9 - - cepstral with dynamic programming
- 10 - - cepstral with heuristics
- 11 - - cepstral with nonlinear smoothing
- 12 - - cepstral with 3 pt median smoothing
- 13 - - cepstral without postprocessing

Figure 6: Evaluation of postprocessing techniques with function O¹ on weighted objective errors.

Contour	pitch error	V-UV error	UV-V error	Total error
Correlation w/o post processing	.52	1.65	.50	2.67
Correlation with 3 pt median smoothing	.66	1.18	.33	2.17
Correlation with nonlinear smoothing	.94	.93	.30	2.17
Correlation with dynamic programming	.35	.86	.39	1.60
Gold Rabiner w/o post processing	.38	4.48	1.03	5.89
Gold Rabiner with 3 pt median smoothing	.25	4.08	.56	4.90
Gold Rabiner with nonlinear smoothing	.24	3.78	.50	4.53
Gold Rabiner with dynamic programming	.65	1.09	.85	2.59
Cepstral w/o heuristics	.40	1.85	2.29	4.54
Cepstral with heuristics	.34	1.64	1.93	3.91
Cepstral with 3 pt median smoothing	.39	1.62	1.85	3.86
Cepstral with nonlinear smoothing	.40	1.52	1.86	3.77
Cepstral with dynamic programming	.33	1.42	.49	2.25

Figure 7: Evaluation of postprocessing techniques with weighted objective measures.